

# LAHNET: A CONVOLUTIONAL NEURAL NETWORK FUSING LOW- AND HIGH-LEVEL FEATURES FOR AERIAL SCENE CLASSIFICATION

Yuansheng Hua<sup>1,2</sup>, Lichao Mou<sup>1,2</sup>, Xiao Xiang Zhu<sup>1,2</sup>

<sup>1</sup>Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany

<sup>2</sup>Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

## ABSTRACT

In this paper, we proposed an innovative end-to-end convolutional neural network (CNN), which is trained to learn how to fuse multi-level features for aerial scene classification. Instead of using only coarse semantic features as conventional CNNs, we resort to first hierarchically extracting dense high-level features and then element-wise fusing them with low-level features to build a comprehensive feature representation, which contains not only high-level semantic information but also fine-grained low-level details, for scene classification. The network is evaluated on two broadly used aerial scene datasets, UCM and AID. The experimental results indicate that the proposed LAHNet performs superiorly compared to the existing benchmark methods. Furthermore, visualization of the fused features presents an intuitive illustration of the remarkable improvement.

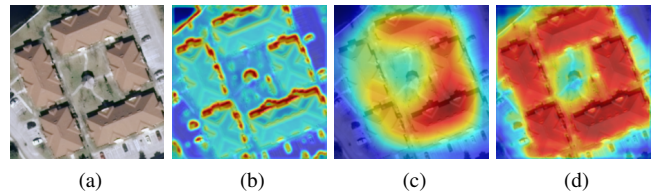
**Index Terms**— convolutional neural network (CNN), feature fusion, aerial scene classification

## 1. INTRODUCTION

With the development of remote sensing technology, an increasing number of high-resolution remote sensing images are now available and widely used in aerial scene classification. However, complicated spatial patterns in high-resolution imagery make identification of aerial scenes a challenging task. To tackle this problem, numerous methods have been proposed during the last decades.

Early studies mainly focus on first extracting low-level visual attributes with feature descriptors, e.g., Scale Invariant Feature Transform (SIFT) descriptors [1], and then encoding these features to build a mid-level representation of aerial scenes [2]. However, these feature representations suffer from high-level semantic information, and their capabilities of identifying scenes of high complexity and small inter-class dissimilarity are limited accordingly.

Recently, convolutional neural network (CNN) based approaches have been proposed and made remarkable success in aerial scene classification [3, 4]. These methods aim at learning high-level semantic features via a hierarchical architecture to predict scene categories. With increasing net-



**Fig. 1:** Comparison of class activation maps (CAMs) [5] of features from different layers. (a) Dense residential (UCM dataset). (b) CAM of low-level features. (c) CAM of high-level features. (d) CAM of fused features in our network.

work depth, more abstract and higher-level features can be extracted. However, the aforementioned methods pay high attention to high-level semantic features, while low-level features, which are rich of fine-grained structure information, are ignored in the final classification.

As a consequence, feature fusion is of high interest, and some pioneer researches have been carried out in this field [6, 7]. Li et al. [6] used Fisher kernel to fuse multilayer features, which are extracted by a pre-trained CNN. Hu et al. [7] proposed a two-stream network for fusing features learned from different data sources. However, in [6], the CNN only serves as a feature extractor, which means the model is not end-to-end, and in [7], multiple data sources are required, which is expensive and costly. Therefore, in this paper, we aim to propose an innovative end-to-end CNN for fusing multi-level features and classifying aerial scenes based on a comprehensive feature representation. Specifically, the network consists of a stack of convolutional layers and atrous convolutional layers, which are utilized to hierarchically extract low-level features and dense high-level features. Afterwards, the extracted high-level feature maps are upsampled and merged with low-level feature maps by element-wise addition to generate final feature representations for scene classification. Besides, the fine-tuned network can be modified to visualize the fused feature maps. Experimental results on two aerial scene datasets demonstrate that the proposed network achieves considerable improvements, and the visualization of fused feature maps show a good balance of both low- and

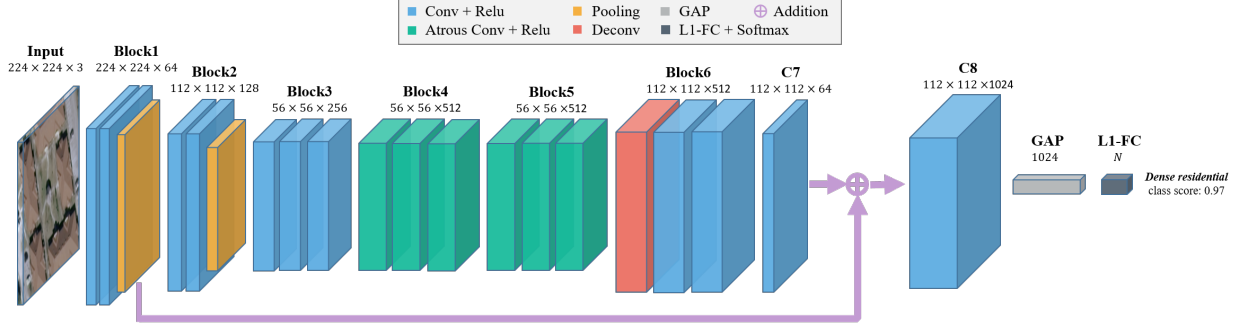


Fig. 2: The architecture of LAHNet.

high-level features.

## 2. NETWORK ARCHITECTURE

Commonly, a CNN employs a set of convolutional layers and pooling layers for hierarchically extracting high-level features. However, with increasing network depth, the extracted high-level feature maps are with a spatially coarse resolution, which is irreversible and not favorable for fusing with low-level feature maps of finer resolution. In this section, we present our network architecture for extracting hierarchical features and a method for fusing multi-level features.

### 2.1. Multi-level feature extraction

Unlike the conventional CNNs, e.g., VGGNet [8], the feature extraction architecture in LAHNet consists of two stages, one convolutional stage and one atrous convolutional stage, as shown in Fig. 2. In the convolutional stage, which consists of Block1, Block2 and Block3, convolutional layers and max-pooling layers are stacked for extracting low-level features and reducing the size of output feature maps, respectively. The receptive field of all convolutional filters is  $3 \times 3$ , and the convolution stride is fixed to 1 pixel. Besides, the pooling window of max-pooling layers is  $2 \times 2$  and the stride is set as 2 pixels, which halves height, and width of output feature maps. It is notable that in the last convolutional block (cf. Block3 in Fig. 2), pooling layer is removed for preserving spatial resolution of the extracted feature maps.

Following the convolutional stage, the atrous convolutional stage, which is only composed of atrous convolutional layers (i.e., Block4 and Block5), is attached for extracting dense high-level features. The motivations of utilizing atrous convolution are as follows: 1) atrous convolutional filters are equipped with enlarged receptive fields, as shown in Fig. 3, which can extract more abstract and holistic features without reducing spatial dimensionality of feature maps or increasing the number of filter parameters and 2) by judiciously designing dilation rate of atrous convolutional filters, e.g., 2 in Block4 and 4 in Block5, it is feasible to initialize LAH-

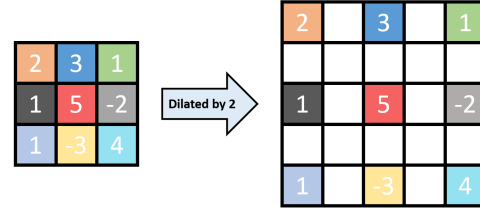


Fig. 3: Comparison between convolution filter (left) and atrous convolution filter (right) with respect to equivalent filter parameters. The dilation rate is 2, and blank cells represent zero.

Net with pre-trained CNN models, e.g., VGGNet, considering that all filters have equivalent receptive fields. Consequently, the extracted high-level feature maps are spatially finer, so called 'dense', compared to feature maps learned from the last convolutional block of conventional CNNs. Moreover, removal of pooling layers guarantees that the spatial dimensionality of final output feature maps is consistent with that of input feature maps.

### 2.2. Fusion of low- and high-level features

To fuse low-level and high-level features, we first upsample high-level feature maps via deconvolution, and then add them to low-level feature maps by an element-wise operation, which are identically mapped from shallow layers via a skip connection, as shown in Fig. 2. The reason of identically mapping low-level features, instead of high-level features, is that the spatial information and structural details learned from shallow layers is vanished with the layer going deep, which means it is impossible to directly generate fine-grained semantic feature maps without any complementary data, e.g., pixel-level labeled training samples [9]. Notably, to reduce the aliasing effect of upsampling and element-wise addition, we make use of additional convolutional layers on upsampled and fused feature maps (cf. Fig. 2).

The size of deconvolutional filters is  $4 \times 4$ , and the stride is 2 pixels for upsampling high-level feature maps. The size

of convolutional filters in C7 is  $1 \times 1$  for reducing channel dimension of feature maps, and the number of filters in C8 is 1024. Furthermore, a global average pooling layer (GAP) is attached, and L1-regularization is applied in the softmax layer for sparsely coding fused feature maps.

### 3. EXPERIMENTS AND DISCUSSION

#### 3.1. Data description

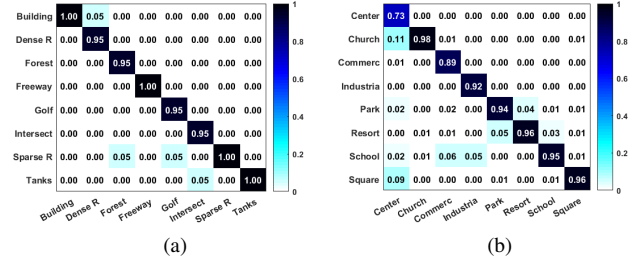
The first dataset, UC-Merced (UCM) dataset [2], consists of 2100 images of  $256 \times 256$  pixels, which are labeled into 21 land-use classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. For each of the 21 classes, 100 images with a spatial resolution of one foot are collected by cropping from large aerial ortho imagery downloaded from the United States Geological Survey (USGS) National Map manually.

The second experimental dataset, Aerial Image Dataset (AID) [10], is a new large-scale image dataset, composed of images of 30 aerial scene classes: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct. All images are collected from world-wide Google Earth imagery and the pixel resolution of AID varies from 8 meters to about half a meter. Moreover, the number of sample images in each category are between 220 to 420, and the entire dataset contains 10000 images of  $600 \times 600$  pixels in total.

#### 3.2. Training details

The proposed LAHNet is initialized with pre-trained VGG-16 model and fine-tuned on UCM and AID datasets, where 80% and 50% of sample images are selected for training, respectively, and the rest for testing. The parameter of L1-regularization is 0.1, and learning rate is initially 0.001 and decayed when accuracy is saturated. The optimizer is set as Nesterov Adam optimizer, and loss function is categorical crossentropy.

We implemented our model on TensorFlow and trained the network on one NVIDIA TITAN X (Pascal) 12GB GPU. The size of training batch is 8 as a compromise between GPU memory capacity and training speed. To avoid overfitting, we stopped training process when validation accuracy decreases after five epochs.



**Fig. 4:** Confusion matrices obtained by LAHNet on (a) UCM and (b) AID datasets.

**Table 1:** The Overall Accuracies (%) of Different Methods

Methods	UCM	AID
BoVW [2]	77.65	68.77
VGG-VD-16 [8]	96.41	90.00
GoogLeNet [11]	95.20	86.94
proposed LAHNet	<b>99.10</b>	<b>95.78</b>

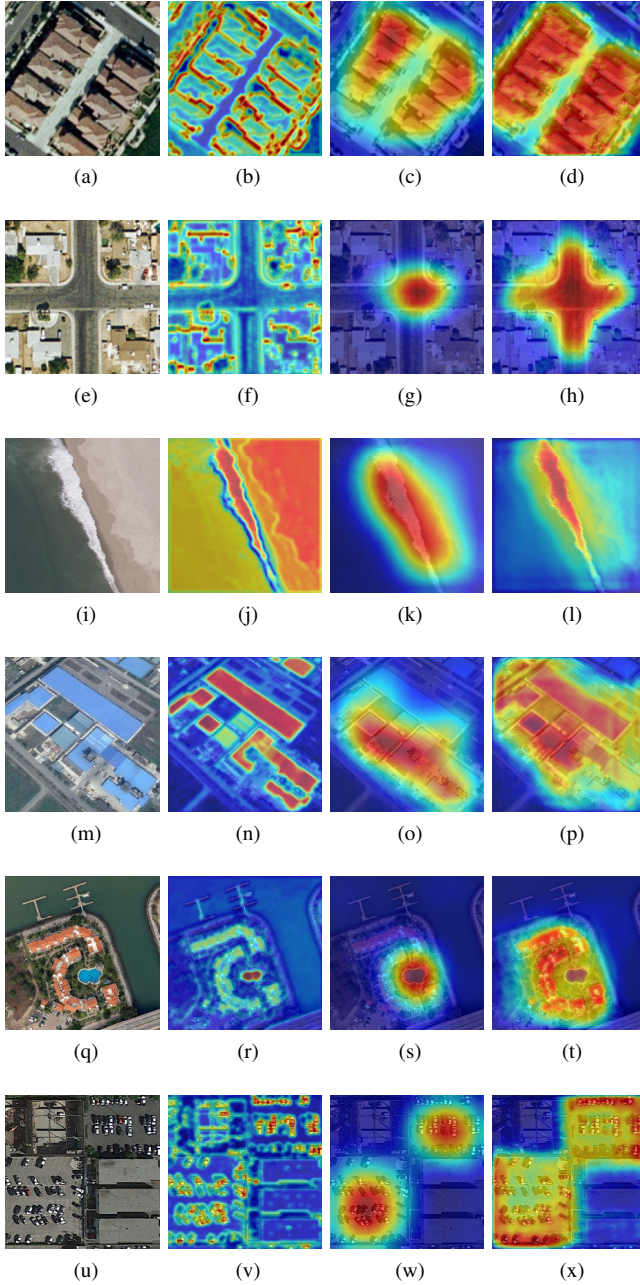
#### 3.3. Discussion of the result

To evaluate the performance of the proposed network, we compute overall accuracy (OA) and compare with benchmark methods, as shown in Table 1. The highest accuracy is marked in bold, and the comparison indicates that our LAHNet outperformed benchmark methods by at least 2.69% on UCM dataset and 5.78% on AID dataset, respectively.

Besides OA, we also provide confusion matrix (cf. Fig. 4) for evaluating the performance on individual classes. Notably, only confusable classes are compared in Fig. 4. According to these quantities, it is evident that distinguishing categories of small inter-class dissimilarity (e.g., "buildings" vs. "dense residential" in the UCM dataset, and "center" vs. "church" in the AID dataset) is not a trivial task. The reason is that both "buildings" and "dense residential" are composed of regular-shaped, man-made constructions, and layouts of the architectures are similar in both categories. For the "center" and "church", it is not difficult to find that roofs of some churches are highly similar to centers from a nadir view, which confuses the network.

Furthermore, to give an insight view of fused features, we generated class activation maps (CAMs) for all sample images, and some examples are shown in Fig. 5. CAMs generated from shallow layer present an explicit view of low-level features, e.g., edges and textures, while CAMs from deep layer highlight coarse discriminative regions. On contrast, CAMs from the fused layer in our network draw a holistic picture of not only where discriminative regions are, but also how the regions appear in detail.





**Fig. 5:** Samples from the UCM and AID datasets. The four columns from left to right are original images, CAMs of low-level features, high-level features, and fused features. In the first column, categories of six samples are (a) dense residential, (e) intersection, (i) beach, (m) industrial, (q) resort and (u) parking lot.

#### 4. CONCLUSION AND OUTLOOK

In this paper, we proposed a novel end-to-end CNN, LAHNet, to fuse multi-level features for aerial scene classification. The network mainly consists of one convolutional stage

and one atrous convolutional stage, which are utilized to hierarchically extract low-level features and dense high-level features. Afterwards, low-level feature maps are identically mapped from shallow layers via a skip connection, and then element-wise added to upsampled high-level feature maps. The experimental results demonstrate that our LAHNet performs superiorly compared to benchmark methods, and fused feature maps are proved to be fine-grained and semantic by their CAMs. Further work mainly comprises extraction of finer semantic features for classification and segmentation.

#### 5. ACKNOWLEDGEMENTS

This work is jointly supported by the China Scholarship Council, the Helmholtz Association under the framework of the Young Investigators Group SiPEO (VH-NG-1018, [www.sipeo.bgu.tum.de](http://www.sipeo.bgu.tum.de)), and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. ERC-2016-StG-714087, Acronym: *So2Sat*).

#### 6. REFERENCES

- [1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [2] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*. ACM, 2010, pp. 270–279.
- [3] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: a review,” *arXiv preprint arXiv:1710.03959*, 2017.
- [4] F. Hu, G. Xia, J. Hu, and L. Zhang, “Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery,” *Remote Sensing*, vol. 7, no. 11, pp. 14680–14707, 2015.
- [5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [6] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, “Integrating multilayer features of convolutional neural networks for remote sensing scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, 2017.
- [7] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, “Fusionet: A two-stream convolutional neural network for urban scene classification using polar and hyperspectral data,” in *Urban Remote Sensing Event (JURSE), 2017 Joint. IEEE*, 2017, pp. 1–4.
- [8] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [9] L. Mou, P. Ghamisi, and X. X. Zhu, “Unsupervised spectral-spatial feature learning via deep residual conv-deconv network for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, 2017.
- [10] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “Aid: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, 2017.
- [11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.